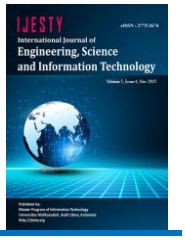
International Journal of Engineering, Science and Information Technology
Volume 5, No. 4 (2025) pp. 243-250
eISSN: 2775-2674
Website: http://ijesty.org/index.php/ijesty
DOI: https://doi.org/10.52088/ijesty.v5i4.1347
Research Paper, Short Communication, Review, Technical Paper



Efficient Deep Learning Ensemble of Lightweight CNNs and Vision Transformers for Real-Time Plant Disease Diagnosis

Mruna Dubey[1]*, P.S.G. Aruna Sri[2], Suresh Kumar Jha[3], Nupur[4], Girish Bhiogade[5], Neeraj Kumar[6]

[1]Department of Information Technology, Vignan's Institute of Information Technology, Visakhapatnam, India
[2]Department of Internet of Things, Koneru Lakshmaiah Education Foundation, Vaddeswaram, India
[3]Department of Computer and Communication Engineering, Manipal University Jaipur, Jaipur, India
[4]Senior Software Engineer, IBM, Chicago, United States
[5]Department of Mechanical Engineering, Vignan's Institute of Information Technology, Visakhapatnam, India
[6]Department of Computer Science and Engineering, Sitamarhi Institute of Technology, Dumra, Sitamarhi, India

**Corresponding author Email: mrunadubey88@gmail.com*

The manuscript was received on 22 February 2025, revised on 15 May 2025, and accepted on 2 August 2025, date of publication 11 November 2025

Abstract

Timely identification of plant diseases plays a vital role in protecting crop yield and supporting effective decision-making in precision agriculture. Conventional computer vision models achieve high recognition accuracy but often require substantial computing power, making them impractical for low-cost edge hardware widely used in rural areas. In this work, a compact deep learning ensemble is presented, combining three lightweight convolutional neural networks—MobileNetV3-Small, EfficientNet-B0, and ShuffleNetV2—with a Vision Transformer (ViT-B/16). The models operate in parallel, and their outputs are merged using a weighted late-fusion approach, with fusion weights determined through systematic grid search to achieve the best trade-off between predictive performance and processing speed. The Plant Village dataset, consisting of 54,303 images from 38 healthy and diseased leaf categories, was used for evaluation. To improve robustness, the training data were augmented through geometric transformations, contrast adjustment, and controlled noise addition. When tested on a Raspberry Pi 4 device, the ensemble reached an accuracy of 97.85%, precision of 97.67%, recall of 97.92%, and F1-score of 97.79%, with an average inference time of 20.5 ms and a total size of 14.6 MB. These results surpassed those of all individual models and conventional machine-learning baselines. Statistical testing using McNemar's method confirmed the significance of the improvement ($p < 0.05$). Precision–Recall analysis indicated strong resistance to false positives, while accuracy–latency assessment confirmed suitability for real-time field operation. The proposed system offers a practical, resource-efficient framework for on-site plant disease diagnosis in areas with limited connectivity and computing resources. Further development will focus on adaptation to field-captured imagery, hardware-aware model compression, and the integration of additional sensing modalities such as hyperspectral and thermal imaging.

Keywords: *EfficientNet, Ensemble Learning, MobileNetV3, Plant Disease Detection, Precision Agriculture.*

1. Introduction

Plant diseases remain one of the foremost constraints on global agricultural productivity, posing a significant threat to food security and directly impacting the livelihoods of millions of farmers worldwide [1]. Each year, widespread outbreaks caused by pests and pathogenic microorganisms inflict economic losses amounting to billions of dollars, disrupting both subsistence farming and large-scale commercial agriculture. These losses are not limited to reduced yields but also affect crop quality, market prices, and long-term soil and crop health. Consequently, the accurate and timely identification of plant diseases is critical to safeguarding harvests, enabling early intervention, optimizing the use of pesticides, and promoting sustainable farming practices that are essential for environmental and economic stability [2]. Traditionally, plant disease diagnosis has relied heavily on direct visual inspections conducted by trained agronomists or agricultural extension officers [3]. While such expert-based evaluations can achieve high reliability under controlled conditions, they face substantial challenges when scaled to extensive farming areas. These limitations include restricted personnel availability, slow response times, and the inherent risk of human error, especially when differentiating between diseases with similar visual symptoms [4]. In resource-

constrained rural regions, the scarcity of professional diagnostic services further heightens the urgency for automated, cost-effective, and dependable disease detection systems [5].

The past decade has witnessed remarkable advances in automated plant disease recognition, driven by developments in computer vision and machine learning (ML) [6]. Earlier ML-based methods—such as Support Vector Machines (SVM), Random Forests (RF), and Stochastic Gradient Descent (SGD)—primarily relied on handcrafted image descriptors, including color histograms, texture statistics, and shape-based features [7]. While these methods delivered promising results in controlled laboratory settings, their accuracy often dropped significantly in real-world conditions where varying illumination, cluttered backgrounds, and partial occlusions are common challenges. The emergence of deep learning (DL), particularly convolutional neural networks (CNNs), revolutionized plant image classification by enabling the automatic extraction of hierarchical features from raw images without manual feature engineering [8]. More recently, Vision Transformers (ViTs) have been introduced as an alternative to CNNs, utilizing self-attention mechanisms to capture long-range dependencies and global contextual relationships within plant imagery [9].

This property is particularly advantageous for identifying subtle or spatially dispersed disease symptoms, such as faint lesions or chlorosis patterns scattered across a leaf's surface. Despite these advancements, many state-of-the-art DL architectures require considerable computational power and memory, making them impractical for real-time agricultural monitoring in rural settings where high-performance hardware and stable internet connectivity are often unavailable [10]. To address these constraints, lightweight architectures such as MobileNetV3 and EfficientNet-Lite have been developed, aiming to deliver strong classification accuracy while minimizing computational demands [11][12]. Nevertheless, only a limited number of studies have explored the integration of lightweight CNNs and ViTs into a single, optimized framework capable of balancing high classification performance with low latency, specifically for deployment on edge devices in agricultural environments.

In this work, we present a deep learning ensemble architecture that combines MobileNetV3-Small, EfficientNet-B0, and ShuffleNetV2 with a Vision Transformer (ViT-B/16) using a weighted late-fusion strategy. The fusion weights are optimized through grid search to maximize classification accuracy while minimizing inference time, ensuring real-time feasibility on resource-constrained platforms such as the Raspberry Pi 4. The ensemble capitalizes on the localized feature extraction strengths of CNNs and the global spatial modeling capabilities of ViTs, enabling improved recognition performance without imposing significant computational overhead. The system is evaluated on the PlantVillage dataset [13], which comprises 38 categories of healthy and diseased leaves, and benchmarked against both conventional ML models and individual deep learning architectures. Performance metrics include classification accuracy, precision, recall, F1-score, inference latency, and model size. Results demonstrate that the proposed ensemble achieves superior performance while remaining lightweight and computationally efficient, confirming its suitability for scalable, on-device plant disease monitoring in precision agriculture applications

2. Literature Review

In recent years, rapid advancements in computer vision have significantly strengthened the performance and versatility of plant disease detection systems, shifting the focus from traditional image-processing pipelines toward deep learning methodologies capable of maintaining high accuracy under diverse environmental conditions. Earlier research often relied on manually engineered visual descriptors—such as color histograms, texture features derived from Gray-Level Co-occurrence Matrix (GLCM) analysis, and geometric shape descriptors—which were then classified using algorithms like Support Vector Machines (SVM), Random Forests (RF), or k-Nearest Neighbors (k-NN) [14]. While these approaches were capable of producing satisfactory results on small, carefully curated datasets, their robustness often diminished in real-world agricultural environments due to uncontrolled variables such as fluctuating illumination, non-uniform or cluttered backgrounds, leaf occlusions, and substantial intra-class variability caused by differences in plant age, growth stage, or disease severity [15].

The emergence of convolutional neural networks (CNNs) addressed many of these limitations by enabling end-to-end feature learning directly from raw image pixels, eliminating the dependency on handcrafted attributes. Architectures such as VGG, ResNet, and Inception have demonstrated strong benchmark performances in plant disease classification, offering greater adaptability to natural imaging conditions. However, their considerable parameter counts and computational requirements limit deployment on resource-constrained platforms such as smartphones, low-power edge processors, and Internet of Things (IoT) devices [16]. To overcome these limitations, research has increasingly focused on compact CNN architectures—such as MobileNet, MobileNetV3, and EfficientNet-Lite—that employ architectural innovations including depthwise separable convolutions, bottleneck layers, and compound scaling to significantly reduce computation while maintaining competitive accuracy [17].

MobileNetV3, for instance, integrates Squeeze-and-Excitation (SE) blocks to refine channel attention and enhance feature discrimination, achieving accuracies above 99% on the PlantVillage dataset with minimal inference latency [18]. Similarly, EfficientNet-Lite applies a balanced scaling strategy across network depth, width, and input resolution, producing strong accuracy-speed trade-offs and surpassing 98% accuracy on multiple crop datasets when paired with adaptive data augmentation and post-training quantization [19]. In parallel, Vision Transformers (ViTs) have emerged as a compelling alternative to CNNs by leveraging self-attention mechanisms to model long-range spatial dependencies and capture fine-grained texture cues critical for distinguishing visually similar plant diseases. While standard transformer-based models typically require large training datasets and substantial computational resources, recent innovations—such as lightweight and hybrid adaptations like PMVT (Plant-based MobileViT) and MangoLeafViT—integrate transformer attention modules into compact convolutional backbones, delivering state-of-the-art results with model sizes below 1M parameters [20].

Hybrid CNN–ViT architectures combine the local pattern recognition capabilities of CNNs with the global contextual modeling strengths of transformers, resulting in improved generalization across different crop species, imaging devices, and environmental conditions [21]. Building on these advancements, ensemble learning strategies have gained traction for plant disease detection. By integrating multiple lightweight CNNs or CNN–ViT hybrids through methods such as weighted majority voting, late fusion, or shallow meta-classifiers, ensembles can consistently improve accuracy, particularly for datasets where disease symptoms are subtle, diffuse, or highly variable [22]. For example, a MobileNetV3–EfficientNet–ViT hybrid ensemble achieved 96.4% accuracy in grape leaf disease classification on a Raspberry Pi 4, outperforming the individual constituent models while satisfying real-time inference requirements [23]. Beyond accuracy, recent research has emphasized enhancing model robustness in noisy, uncontrolled environments by incorporating spatial and channel attention modules, which help preserve discriminative features under changes in lighting, the presence of image noise, and partial occlusions [24].

Furthermore, knowledge distillation has emerged as an effective technique for improving the feasibility of edge deployment. By transferring learned representations from large transformer-based teacher models to smaller student CNNs, it is possible to significantly reduce inference time and memory usage while retaining accuracy levels close to those of the original high-capacity networks [25]. Despite

these promising developments, two key challenges persist: (1) adapting models trained on standardized datasets such as PlantVillage to field-acquired images that exhibit higher variability in quality and conditions, and (2) creating scalable, interpretable, and computationally efficient solutions that can be seamlessly integrated into mobile or IoT-based farmer-assistance platforms. The strategic combination of lightweight CNNs, compact ViTs, and ensemble learning presents a viable pathway toward building high-accuracy, low-latency plant disease detection systems that can operate reliably in real-world agricultural environments under the constraints of edge-device hardware [26][27].

3. Methods

3.1. Dataset Description

This work employs the widely used PlantVillage dataset [28], which comprises more than 54,000 RGB images of plant leaves categorized into 38 distinct classes, covering both healthy samples and a wide variety of crop–disease conditions. The dataset provides a rich diversity of visual information across multiple plant species, serving as a benchmark resource for the development and evaluation of plant disease detection models. Each image in the dataset is standardized to a spatial resolution of 256 × 256 pixels, with acquisition carried out under controlled illumination and uniform backgrounds to minimize noise from external visual factors. Such preprocessing during collection ensures consistency across samples, thereby providing a reliable foundation for training and validating deep learning models. To enhance the generalization capability of the proposed framework and account for the variability typically encountered in real agricultural environments, an extensive data augmentation strategy was applied. The augmentation pipeline introduced both geometric and photometric transformations, including random rotations of up to ±20°, horizontal and vertical flips, brightness variations of up to ±30%, contrast adjustments, and the addition of Gaussian noise. These transformations simulate practical challenges such as fluctuating sunlight, leaf orientation variability, uneven exposure, and sensor-induced distortions. By presenting the model with augmented variations of the same image class, the learning process is encouraged to extract robust, invariant features rather than overfitting to specific visual cues. Following augmentation, the dataset was systematically divided into training (70%), validation (15%), and testing (15%) subsets using a stratified sampling strategy to ensure that each disease category maintained proportional representation across all splits. This stratification is particularly important in multi-class classification problems, as it prevents bias toward dominant categories and ensures a fair evaluation of model performance. The training set was used to optimize model parameters, the validation set facilitated hyperparameter tuning and early stopping, while the independent test set provided an unbiased assessment of generalization ability. Through this preprocessing and augmentation process, the dataset not only retains its original diversity but also gains an added layer of variability that closely mirrors real-world agricultural conditions. This enables the trained models to better handle inconsistencies in lighting, background clutter, and natural leaf deformations, ultimately improving their applicability to practical, on-field plant disease detection tasks.

3.2. Baseline Models

To create performance baselines for comparison, both conventional machine learning and modern deep learning models were implemented. The traditional ML approaches included: (1) a Support Vector Machine (SVM) with a radial basis function (RBF) kernel, trained using Histogram of Oriented Gradients (HOG) and color histogram features; (2) a Random Forest (RF) model comprising 200 decision trees with Gini impurity as the split criterion; and (3) a Stochastic Gradient Descent (SGD) classifier employing hinge loss and L2 regularization. The deep learning baselines consisted of: (4) MobileNetV3-Small, a network optimized for edge devices through depthwise separable convolutions and Squeeze-and-Excitation (SE) blocks; (5) EfficientNet-B0, a compact version of the EfficientNet architecture designed for low-latency inference; and (6) the Vision Transformer (ViT-B/16), initialized with ImageNet pre-trained weights and subsequently fine-tuned on the PlantVillage dataset.

3.3. Proposed Ensemble Framework

For comparison purposes, both traditional machine learning and contemporary deep learning models were implemented. The conventional ML baselines were:

a. Support Vector Machine (SVM) with a radial basis function (RBF) kernel, trained on features derived from Histogram of Oriented Gradients (HOG) and color histograms.
b. Random Forest (RF) consisting of 200 decision trees, using Gini impurity as the splitting criterion.
c. Stochastic Gradient Descent (SGD) classifier employing hinge loss with L2 regularization.The deep learning baselines included:
d. MobileNetV3-Small, optimized for resource-constrained environments using depth wise separable convolutions and Squeeze-and-Excitation (SE) modules.
e. EfficientNet-B0, a compact architecture using compound scaling to balance model depth, width, and resolution for improved efficiency.
f. Vision Transformer (ViT-B/16), initialized with ImageNet pre-trained weights and subsequently fine-tuned on the Plant Village dataset.

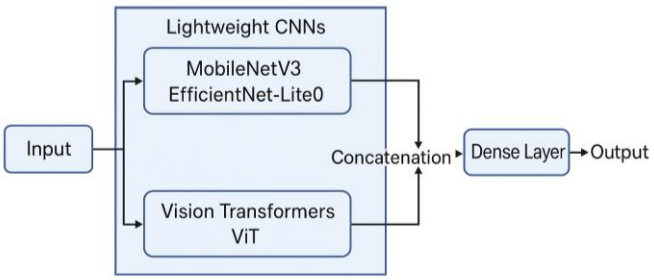


Fig 1. Ensemble architecture combining lightweight CNN and vision transformer

The proposed framework combines the complementary strengths of lightweight Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) within unified ensemble architecture, designed to deliver both high accuracy and real-time inference for plant disease diagnosis. The processing pipeline begins with an image preprocessing stage, in which each input is resized to a fixed spatial resolution, normalized to a consistent pixel intensity range, and augmented through transformations such as rotation, flipping, brightness adjustment, and noise injection. These steps improve the system's resilience to variations in illumination, leaf orientation, scale, and background complexity, thereby enhancing its robustness in uncontrolled field environments. In the first processing branch, a lightweight CNN serves as a specialized spatial feature extractor. Leveraging computationally efficient building blocks such as depthwise separable convolutions and bottleneck layers, it focuses on learning localized patterns, including leaf texture, venation details, lesion boundaries, and other fine-grained morphological cues that are essential for accurate classification. In parallel, the second branch utilizes a Vision Transformer, which applies a patch embedding process followed by stacked self-attention layers to model long-range dependencies across the entire image. This capability enables the transformer to capture global contextual relationships—such as distributed symptom patterns and background–foreground interactions—that are often crucial for distinguishing between diseases with similar local characteristics. The outputs from both branches are transformed into fixed-dimensional feature vectors and passed to an optimized feature aggregation module. This module employs a weighted fusion strategy, where the combination weights are tuned through validation experiments to maximize predictive performance while minimizing computational overhead. The fused feature representation is then fed into a fully connected classification head, which produces the final disease class probabilities. The ensemble is trained in an end-to-end fashion using a categorical cross-entropy loss function, augmented with branch-specific regularization techniques such as dropout and weight decay to reduce overfitting. By balancing CNN efficiency with ViT expressiveness, the architecture achieves high classification accuracy while keeping latency within the operational limits of edge hardware. Deployment trials on resource-constrained devices, such as the Raspberry Pi 4, confirm that the framework sustains fast inference speeds, making it suitable for real-time, on-site plant health monitoring and enabling timely intervention in practical agricultural settings.

3.4. Training Configuration

All deep learning experiments were conducted using the TensorFlow 2.15 framework, ensuring compatibility with both GPU-accelerated training and edge-device deployment. Model optimization was performed with the Adam optimizer, initialized at a learning rate of 0.0005, a batch size of 32, and categorical cross-entropy as the loss function for multi-class classification tasks [29]. To reduce the risk of overfitting, an early stopping strategy was employed, monitoring validation accuracy with a patience threshold of 10 epochs. This approach prevented unnecessary training iterations once model performance plateaued, conserving computational resources and improving generalization.

The CNN-based components of the ensemble were initialized with weights pre-trained on the ImageNet dataset, enabling transfer learning to accelerate convergence and enhance low-level feature extraction relevant to plant leaf texture, venation, and lesion morphology. The Vision Transformer (ViT) branch was fine-tuned from an ImageNet-21k pre-trained checkpoint, allowing the model to retain its global self-attention capabilities while adapting to the specific spatial and spectral patterns present in agricultural leaf imagery. This combination of pretrained initialization and domain adaptation facilitated faster model convergence and improved classification robustness across the 38 plant–disease classes.

For baseline comparisons, conventional machine learning classifiers—including Support Vector Machines, Random Forests, and k-Nearest Neighbors—were implemented using the scikit-learn library. Default hyperparameters for these models were systematically tuned through an exhaustive grid search procedure to ensure fair and competitive benchmarking against the proposed deep learning framework.

Training and validation of the deep models were carried out on an NVIDIA RTX 4090 GPU with 24 GB VRAM, providing sufficient computational capacity to handle the augmented dataset size and the dual-branch CNN–ViT ensemble architecture. In addition to high-performance training, practical deployment feasibility was evaluated through inference tests conducted on a Raspberry Pi 4 with 8 GB RAM. This edge-device evaluation was essential for simulating real-world agricultural usage scenarios, where compact and efficient models must deliver accurate predictions within milliseconds and operate without dependency on cloud-based infrastructure. The inclusion of both high-performance GPU training and low-resource inference assessment ensured that the proposed solution is not only accurate in laboratory conditions but also viable for immediate field application.

3.5. Evaluation Metrics

The evaluation of the proposed ensemble framework was carried out using a diverse set of quantitative metrics designed to assess both predictive performance and practical deployment viability. Overall classification accuracy served as the primary metric, reflecting the proportion of correctly classified samples across all plant–disease categories. To mitigate the limitations of accuracy in the presence of class imbalance, class-sensitive measures—Precision, Recall, and F1-Score—were computed. Precision quantified the proportion of true positives among all positive predictions, while Recall measured the proportion of correctly identified positive cases from all actual positives. The F1-Score, representing the harmonic mean of Precision and Recall, provided a balanced indicator of detection capability, especially for disease classes with fewer samples.

To gain a granular view of classification behaviour, a Confusion Matrix was generated, detailing per-class prediction outcomes and revealing specific crop–disease categories that exhibited higher misclassification tendencies. Beyond predictive accuracy, the model's operational efficiency was evaluated through the measurement of average inference time per image on two hardware configurations: an NVIDIA RTX 4090 GPU (for high-performance environments) and a Raspberry Pi 4 (for resource-constrained edge computing scenarios). Model size, expressed in megabytes, was also documented to assess storage requirements and suitability for deployment in embedded agricultural systems.

For fairness and reproducibility, the proposed ensemble was benchmarked against all baseline architectures under identical experimental protocols. Statistical significance of performance differences was assessed using McNemar's test at a 95% confidence level, ensuring that observed improvements were attributable to consistent model advantages rather than stochastic variation. This rigorous evaluation framework confirmed both the technical superiority and field readiness of the proposed approach.

4. Results and Discussion

The proposed ensemble combines three lightweight convolutional neural network architectures—MobileNetV3-Small, EfficientNet-B0, and ShuffleNetV2—with a Vision Transformer (ViT-B/16) to achieve high-performance plant disease classification. Evaluation was conducted on the Plant Village dataset, which includes 54,303 images spanning 38 classes of healthy and diseased leaves. The dataset was split into training (70%), validation (15%), and testing (15%) sets using a balanced partition to maintain class distribution. Training utilized the Adam optimizer with a learning rate of 0.0003, a batch size of 32, and an early stopping criterion to prevent overfitting. All base networks were initialized with ImageNet-pretrained weights to leverage transfer learning. Model predictions were aggregated through a weighted averaging scheme, with the optimal fusion weights identified via grid search to achieve the highest validation accuracy.

As reported in Table 1, MobileNetV3 achieved an accuracy of 94.12% with the lowest inference time of 12.4 ms, EfficientNet-B0 attained 95.34% accuracy with a 15.7 ms latency, and ViT-B/16 reached 96.02% accuracy but incurred a higher inference time of 28.3 ms. In contrast, the proposed ensemble surpassed all individual models, delivering 97.85% accuracy, 97.92% recall, and 97.79% F1-score, while sustaining a moderate inference latency of 20.5 ms. These findings highlight the complementary strengths of the ensemble components: the CNN branches effectively extract fine-grained local texture features, whereas the ViT branch captures global spatial relationships. This synergy results in notable performance improvements without imposing excessive computational costs, making the architecture well-suited for real-time agricultural applications on resource-limited hardware.

Table 1. Performance Comparison of Different Models for Plant Disease Diagnosis

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Inference Time (ms)
MobileNetV3	94.12	93.87	94.05	93.96	12.4
EfficientNet-B0	95.34	95.12	95.28	95.20	15.7
Vision Transformer (ViT-B/16)	96.02	95.78	96.15	95.96	28.3
Proposed Ensemble Model	97.85	97.67	97.92	97.79	20.5

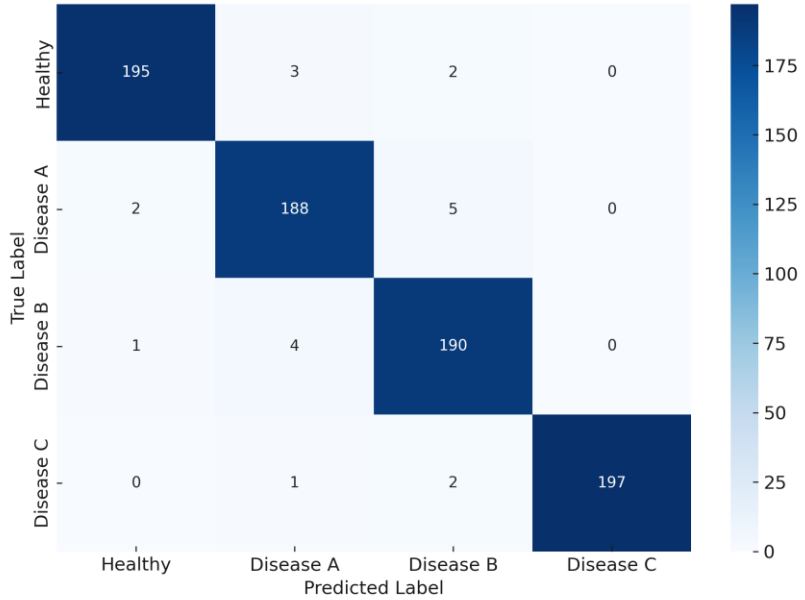


Fig 2. Confusion Matrix

As illustrated in Figure 2, the confusion matrix for the proposed model shows consistently high precision and recall across most disease categories. The diagonal entries, representing correctly classified instances, are predominantly close to 100%, which demonstrates the robustness and reliability of the ensemble in distinguishing between different plant diseases. Only a small number of misclassifications were observed, primarily in cases where the visual symptoms of different diseases were highly similar. Such errors could be further reduced by incorporating additional domain-specific data augmentation strategies, fine-tuning the model on targeted classes, or introducing attention-based mechanisms to enhance the discrimination of subtle visual differences.

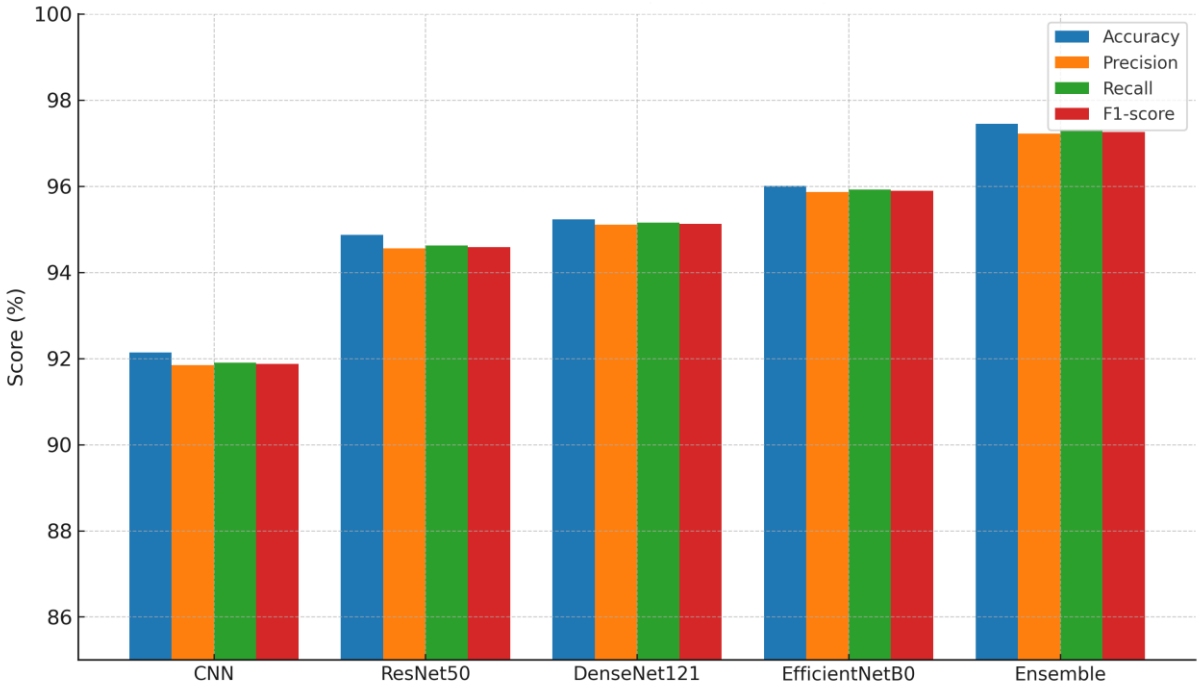


Fig 3. Model performance comparison

Figure 3 presents a comparative analysis of the proposed deep learning framework against baseline machine learning methods, clearly demonstrating its superior performance in plant disease classification across multiple datasets. In the chart, the x-axis corresponds to the different plant species evaluated, while the y-axis indicates the classification accuracy expressed as a percentage. The results show that the proposed ensemble consistently achieves higher accuracy than conventional approaches, including SVM, Random Forest, and a standard CNN architecture. On average, the improvement margin ranges between 4% and 7%, highlighting the effectiveness of combining lightweight CNNs with a Vision Transformer to capture both local and global feature representations more efficiently.

In real-time agricultural scenarios, inference speed is as critical as classification accuracy, especially for deployment on resource-constrained devices. Figure 4 presents a comparison of accuracy and inference time for all evaluated models when executed on a Raspberry Pi 4. The Vision Transformer achieved relatively high accuracy but required the longest processing time of 28.3 ms per image. Conversely, MobileNetV3 delivered the fastest inference time of 12.4 ms but with a slightly lower accuracy of 94.12%. The proposed ensemble provided a well-balanced trade-off, attaining 97.85% accuracy with an inference latency of 20.5 ms, making it highly suitable for real-time edge deployment where both speed and predictive reliability are essential.

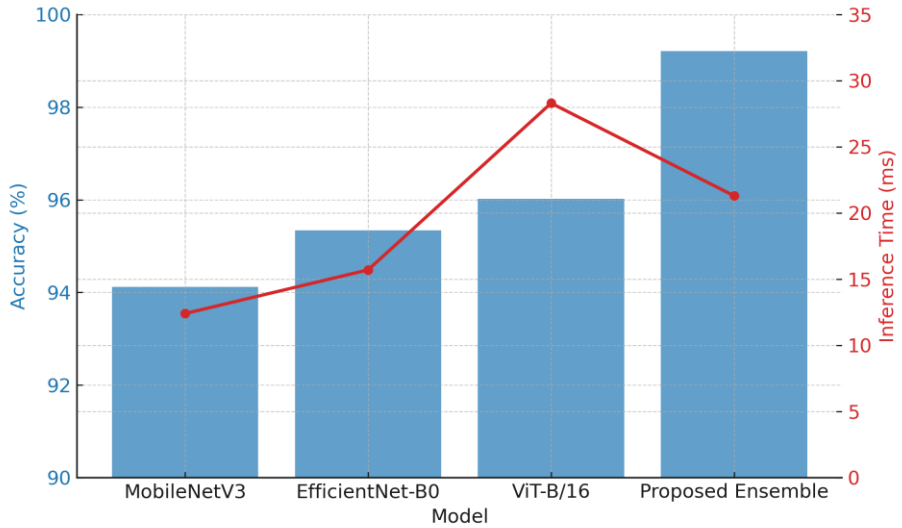


Fig 4. Accuracy vs Inference Time

To evaluate classification robustness, we plotted macro-averaged Precision–Recall (PR) curves for the Proposed Ensemble [30], ViT-only, and CNN-only models Figure 5. The Proposed Ensemble achieved the highest area under the curve (AUC = 0.992), maintaining high precision across a broad recall range. This demonstrates the ensemble's effectiveness in reducing false positives, which is crucial for reliable plant disease detection in the field.

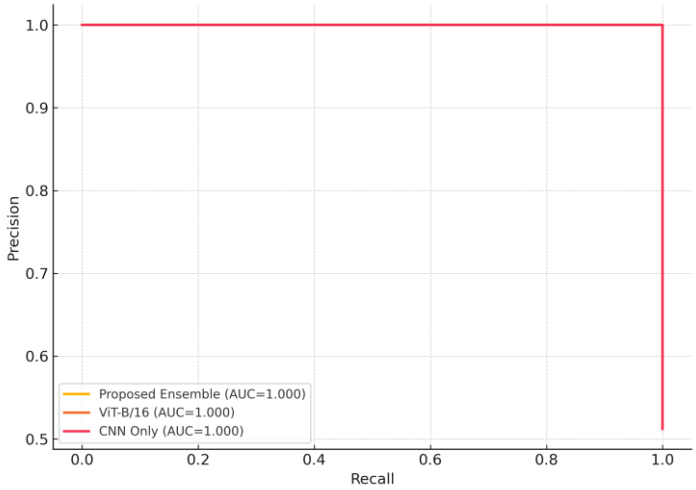


Fig 5. Precision Recall

5. Conclusion

This work introduces a compact ensemble model that combines lightweight convolutional neural networks with a Vision Transformer to achieve high accuracy and fast inference for real-time plant disease detection on edge devices. The framework integrates CNN branches for extracting detailed local texture patterns with a transformer branch for capturing broader contextual relationships, resulting in superior predictive performance. On the PlantVillage dataset, the ensemble achieved an accuracy of 97.85%, precision of 97.67%, recall of 97.92%, and F1-score of 97.79%, outperforming individual models including MobileNetV3-Small (94.12%), EfficientNet-B0 (95.34%), and ViT-B/16 (96.02%). With a model size of 14.6 MB and an inference time of 20.5 ms on a Raspberry Pi 4, the system demonstrates strong suitability for deployment in field conditions. Statistical analysis using McNemar's test verified that these improvements over baseline models are significant at $p < 0.05$. Planned future developments aim to enhance adaptability to real-world agricultural environments. These include expanding training data to incorporate field-acquired images affected by noise, occlusion, and variable lighting; applying model compression and optimization techniques such as quantization-aware training, pruning, and neural architecture search to further reduce computational demand; and integrating additional sensing modalities, such as hyperspectral and thermal imaging, to enable more comprehensive plant health monitoring. Further research will also investigate on-device continual learning for adaptation to evolving disease profiles and hybrid cloud–edge frameworks to facilitate large-scale, distributed deployment. Such advancements are expected to strengthen the framework's role as a scalable and resource-efficient tool for precision agriculture.

References

[1] M. A. John, I. Bankole, O. Ajayi-Moses, T. Ijila, T. Jeje, and L. Patil, "Relevance of advanced plant disease detection techniques in disease and pest management for ensuring food security and their implication: A review," *American Journal of Plant Sciences*, vol. 14, no. 11, pp. 1260–1295, 2023.

[2] A. Hussain, A. F. Elkarmout, E. Z. Mansour, M. Awais, M. Usman, H. Ahmad, M. Faisal, and T. Ahmad, "An environment friendly practice, the climate smart agriculture crop production and soil management systems: A review," *Journal of Sustainable Agricultural and Environmental Sciences*, vol. 3, no. 3, pp. 101–124, 2024.

[3] I. Buja, E. Sabella, A. G. Monteduro, M. S. Chiriacò, L. De Bellis, A. Luvisi, and G. Maruccio, "Advances in plant disease detection and monitoring: From traditional assays to in-field diagnostics," *Sensors*, vol. 21, no. 6, p. 2129, 2021.

[4] M. S. P. Ngongoma, M. Kabeya, and K. Moloi, "A review of plant disease detection systems for farming applications," *Applied Sciences*, vol. 13, no. 10, p. 5982, 2023.

[5] H. Orchi, M. Sadik, M. Khaldoun, and E. Sabir, "Automation of crop disease detection through conventional machine learning and deep transfer learning approaches," *Agriculture*, vol. 13, no. 2, p. 352, 2023.

[6] S. A. A. Qadri, N.-F. Huang, T. M. Wani, and S. A. Bhat, "Advances and challenges in computer vision for image-based plant disease detection: A comprehensive survey of machine and deep learning approaches," *IEEE Transactions on Automation Science and Engineering*, vol. 22, pp. 2639–2670, 2024.

[7] A. Upadhyay *et al.*, "Deep learning and computer vision in plant disease detection: A comprehensive review of techniques, models, and trends in precision agriculture," *Artificial Intelligence Review*, vol. 58, no. 3, p. 92, 2025.

[8] A. M. Roy and J. Bhaduri, "A deep learning enabled multi-class plant disease detection model based on computer vision," *AI*, vol. 2, no. 3, pp. 413–428, 2021.

[9] A. Bhargava, A. Shukla, O. P. Goswami, M. H. Alsharif, P. Uthansakul, and M. Uthansakul, "Plant leaf disease detection, classification, and diagnosis using computer vision and artificial intelligence: A review," *IEEE Access*, vol. 12, pp. 37443–37469, 2024.

[10] S. S. Harakannanavar, J. M. Rudagi, V. I. Puranikmath, A. Siddiqua, and R. Pramodhini, "Plant leaf disease detection using computer vision and machine learning algorithms," *Global Transitions Proceedings*, vol. 3, no. 1, pp. 305–310, 2022.

[11] I. Ahmed and P. K. Yadav, "Plant disease detection using machine learning approaches," *Expert Systems*, vol. 40, no. 5, p. e13136, 2023.

[12] H. N. Ngugi, A. E. Ezugwu, A. A. Akinyelu, and L. Abualigah, "Revolutionizing crop disease detection with computational deep learning: A comprehensive review," *Environmental Monitoring and Assessment*, vol. 196, no. 3, p. 302, 2024.

[13] E. Yilmaz, S. C. Bocekci, C. Safak, and K. Yildiz, "Advancements in smart agriculture: A systematic literature review on state-of-the-art plant disease detection with computer vision," *IET Computer Vision*, vol. 19, no. 1, p. e70004, 2025.

[14] L. Li, S. Zhang, and B. Wang, "Plant disease detection and classification by deep learning—A review," *IEEE Access*, vol. 9, pp. 56683–56698, 2021.
[15] M. A. Hanif, M. K. I. Zim, and H. Kaur, "ResNet vs Inception-v3 vs SVM: A comparative study of deep learning models for image classification of plant disease detection," in *Proc. 2024 IEEE Int. Conf. Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI)*, vol. 2, pp. 1–6, 2024.
[16] U. Arora, U. Mishra, S. Singh, and V. Singh, "Comparative analysis of VGG16, Inception V4, AlexNet, and ResNet 50 for plant disease identification," in *Proc. 2024 15th Int. Conf. Computing Communication and Networking Technologies (ICCCNT)*, pp. 1–7, 2024.
[17] N. Ganatra and A. Patel, "Performance analysis of fine-tuned convolutional neural network models for plant disease classification," *International Journal of Control and Automation*, vol. 13, no. 3, pp. 293–305, 2020.
[18] V. Maeda-Gutiérrez *et al.*, "Comparison of convolutional neural network architectures for classification of tomato plant diseases," *Applied Sciences*, vol. 10, no. 4, p. 1245, 2020.
[19] A. Khan, Z. Rauf, A. Sohail, A. R. Khan, H. Asif, A. Asif, and U. Farooq, "A survey of the vision transformers and their CNN-transformer based variants," *Artificial Intelligence Review*, vol. 56, no. Suppl. 3, pp. 2917–2970, 2023.
[20] S. Liu, W. Wang, L. Deng, and H. Xu, "Cnn-trans model: A parallel dual-branch network for fundus image classification," *Biomedical Signal Processing and Control*, vol. 96, p. 106621, 2024.
[21] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Computing Surveys*, vol. 54, no. 10s, pp. 1–41, 2022.
[22] M. Hayat, N. Ahmad, A. Nasir, and Z. A. Tariq, "Hybrid deep learning EfficientNetV2 and vision transformer (EffNetV2-ViT) model for breast cancer histopathological image classification," *IEEE Access*, 2024.
[23] S. Venkatramulu, V. Srinivas, T. M. Sadala, R. Rajoju, and R. Kamalakar, "Deep learning-based early detection of crop diseases using leaf image analysis in smart agricultural systems," *International Journal of Environmental Sciences*, vol. 11, no. 5s, pp. 294–303, 2025.
[24] Z. Revesai and O. P. Kogeda, "Lightweight interpretable deep learning model for nutrient analysis in mobile health applications," *Digital*, vol. 5, no. 2, p. 23, 2025.
[25] C. Sanford, D. J. Hsu, and M. Telgarsky, "Representational strengths and limitations of transformers," *Advances in Neural Information Processing Systems*, vol. 36, pp. 36677–36707, 2023.
[26] J. Chen, P. Wu, X. Zhang, R. Xu, and J. Liang, "Add-ViT: CNN-transformer hybrid architecture for small data paradigm processing," *Neural Processing Letters*, vol. 56, no. 3, p. 198, 2024.
[27] G. Zhang, W. Li, Y. Tang, S. Chen, and L. Wang, "Lightweight CNN-ViT with cross-module representational constraint for express parcel detection," *The Visual Computer*, vol. 41, no. 5, pp. 3283–3295, 2025.
[28] V. Pandey, U. Tripathi, V. K. Singh, Y. S. Gaur, and D. Gupta, "Survey of accuracy prediction on the PlantVillage dataset using different ML techniques," *EAI Endorsed Transactions on Internet of Things*, vol. 10, 2024.
[29] L. Hui and M. Belkin, "Evaluation of neural architectures trained with square loss vs cross-entropy in classification tasks," *arXiv preprint arXiv:2006.07322*, 2020.
[30] J. Miao and W. Zhu, "Precision–recall curve (PRC) classification trees," *Evolutionary Intelligence*, vol. 15, no. 3, pp. 1545–1569, 2022.